Journal of Theoretical and Applied Information Technology
15th May 2023. Vol.101. No 9
© 2023 Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DETECTING USER'S INTENT BY COMMUNITY QUESTION ANSWERING FOR INFORMATION RETRIEVAL

MEHDI GAOU[1] , HICHAM TRIBAK[2], SALAH KRIT[3], SALMA GAOU[4]

[1]Faculty of science, Laboratoire des Sciences de l'Ingénieur (LabSIE), Ibn Zohr University, Morocco

[234]Polydisciplinary Faculty. Ibn Zohr University, Morocco

E-mail: [1]gaou40@hotmail.com, [2] h.Tribak@uiz.ac.ma, [3]s.krit@uiz.ac.ma, [4]s.gaou@uiz.ac.ma

Abstract

Information retrieval systems (IRs) have seen new types of tools called Community Answering Questions (CQA). It is the taking into account of the need for precise information of the user that motivated the emergence of such systems. A Community Answering Questions (CQA) system can be opposed to an Internet search engine like Google or Yahoo! Wiki Answers, Answers and domain-specific forums like Stack Overflow in certain specific points. Although the idea of receiving a direct and targeted response to an issue seems very attractive, the quality of the question itself can have a significant effect on the likelihood of obtaining useful responses. Such an information retrieval paradigm is particularly appealing when the problem cannot be answered directly by the search engines due to the unavailability of relevant online content. A good understanding of the underlying purpose of an issue is essential to meet the information needs of the user better.

In this article, we analyze the intent of each question in CQA, the research problem arising from the previously stated objective consists in estimating the best answer according to a question, all its responses and the metadata attached to it. The CQA is reducible to a classification problem, with the "best" answers as a particular class, the rest as a negative class. We can obtain significant and significant improvements in classification concerning state of the art in this field. In addition to textual features, a variety of metadata features are used to model a user's intent, which helps the CQA service better answer to similar questions — recommending the most relevant respondents.

Keywords: *Community Question Answering, Question Retrieval, User Intent, Modeling Entry Point Prdiction.*

1. INTRODUCTION

With the bifurcation of the web from a predominantly vertical distribution system (a broadcaster for a multitude of consumers) to a mostly horizontal communication system (each consumer is also a broadcaster), many sites whose sharing and communication between user's heart of the functioning have appeared. These sites, such as social networks (for example, Facebook, LinkedIn), social bookmarking sites (e.g. Connotea1), micro-blogging sites (e.g. Twitter2), constitute what was called by some enthusiasts of the world the "web 2.0 revolution", the "social web" or, more modestly, the "participative web" [8].

From this new wave of sites whose content is generated by its contributors were born Community Question-Answering (CQA) sites. These sites allow users to create questions, answer questions from other users, comment on various issues and answers, and judge the relevance of other users' responses using scoring devices (score of 5, positive vote / Negative, etc.).

Information retrieval systems (IRs) have seen new types of tools called Community Question-Answering systems. It is the taking into account of the need for precise information of the user that motivated the emergence of such systems. A CQA can be opposed to an Internet search engine like Google or Yahoo! On some specific points, the user enters his query in natural language (LN) and not in the form of keywords. Downstream, the system offers the answers expected by the user and does not return a few thousand documents to traverse manually. Two different uses emerge clearly between the two types of tools. While search engines can retrieve documents on a general theme, question answering systems are used to find accurate information, which are a few words.[1]

In CQA, users regularly ask questions in natural language, which are addressed to humans, while in web search users submitting keyword queries that directed to computer algorithms. More precisely, this leads to the following five significant differences between the questions of the CQA and the queries of the search engines:

•Many CQA questions are intrinsically subjective. It has been shown that the proportion of Yahoo! Answers-oriented factual question answering is flipping while personal/ complex problem - Answers is gradually becoming more [17]

• Many CQA questions are socially motivated because users know that the answers to their questions would come from other users in the community. Instead of satisfying a need for information, these questions are in fact about social ties (for example, finding a date), or about the generation of some empathy (e.g. complaining), or only for entertainment purposes (e.g. saying jokes).

• Even though many numbers of queries submitted to search engines are formatted [18], they are quite different from the query patterns used in CQA services. For example, instead of using the common question format "What is a", or "Where is" in CQA, search queries in search engines are more likely to be formatted like "I Need "," I want "," Show me."

• System CQA problems are more likely to have additional constraints because they are usually longer and more complex than search engine queries. For example, people may ask for something in a specific area (e.g. looking for the shop), or within a particular period of time (e.g. looking for information about school).

• Compared to search engines, the CQA services have more precious data, which can be used to characterize a person's social status. For example, each user has their unique request and respond to the story; each question may correspond to a better answer, and /an up-vote down-vote value; besides, some users.

• Have the reason to ask questions in several specific subjects (e.g. Travel).

Our proposal is the realization of an automatic method to find the best entry point for an information retrieval system in a community question-answers system (CQA). A vast amount of information, more than a billion answers for Yahoo! Answers only by [30], is available in CQA systems in a form which is slightly exploited. A user seeking a reply to a question is obliged to see all the answers to all items similar to his or her own.

2. STATE OF THE ART

Community Question- Answering (CQA) systems can be examined from two angles: their organization (thematically), and their ranking of answers (by social relevance).

At the organizational level, CQA systems may be similar to systems(Social Tagging Systems, [6]) in the fact that they use tags, which are necessarily free keywords and a complete study of which can find in [10] Users. These markers can be used as a glossary of the field, because of their tendency to converge towards an almost constant number. Social tagging systems are considered a collective intelligence trace [6].

In the ranking of answers, CQA systems traditionally use a self-organized voting system, as can be seen in Yahoo! Answers, which captures a large proportion of the market with more than a billion questions [30]. Each user can give a positive or negative vote on a contribution (question or answer) that looks differently to allow the system to remove contributions that the community does not appreciate (majority of negative votes) and return to those that the community considers relevant (majority of affirmative votes) This system allows creating a consensual criterion on which to classify the answers: the aggregation of the positive and negative votes, which will be called social relevance score or social score.

Harper et al. In [15] detail the predictor indices of the quality of answers in a CQA system. The notion of answers were not fixed in the system (and therefore not directly extractable), they submitted to a panel of users a set of questions posed by the authors and the answers associated with them in order to obtain a judgment user. The aim of the authors was to discover the differences between question-answers sites and to establish a typology:

- digital reference services, where the system is analogous to a library search service;
- "request to expert" sites, where the system is hierarchical and made up of experts from different fields;

The CQA sites, which are the subject of this paper.

[1]

The authors conclude that if the "request to an expert" pay services produce high-quality answers, the CQA sites are second to very small if the author of the question makes the formulation.

Jurczyk and Agichtein in [1] [2] exploit link analysis techniques such as the HITS algorithm (more details on HITS are available in [11]) to discover the user-authorities. The analogy with the CQA sites is done by considering each question in two dimensions: a bad question does not attract many answers (and therefore has a low outgoing degree), and the opposite for a good question. Users who focus on "good" questions have a high degree of intake. The analogy is thus made with the authors of these questions (which are "hubs") and the authors of the answers on these questions (which are "authorities").

Yang et al. In [9] attempt to predict questions of CQA systems that will not be answered. They use heuristics such as the length of the question or the history of the questioner to train classifiers. The classifiers used are Naive Bayes, a decision tree (J48), AdaBoost and an SVM (driven with the SMO algorithm). The authors also deal with the problem of imbalance in the dataset, due to the fact that the number of questions answering is largely the majority, with a simple resampling. The authors analyze the characteristics used and discuss their discriminating power.

Shah and Pomerantz in [12] discuss quality assessment in a CQA system. The authors focus on a human evaluation based on criteria pre-chosen by them. While it is easy to reach a consensus on the criteria defined by the authors, it seems more difficult to predict a better answer using only these collective judgments. The authors then construct a model based on automatically extracted attributes and reach a precision of 81% after cross-validation at 10 assays.

Jeon et al. In [3] attempt to establish a theoretical framework for predicting the quality of answers using non-textual variables, The authors use a probabilistic formalism in order to be able to reuse the theoretical framework in coupling with other models. In this sense, they are the closest to our work. The predictor thus created favors recall (92% against a precision of 65%).

In [5] Liu et al. Focus on the subjectivity of the notion of quality of answer. To this end, they create their own criterion of "satisfaction", which they consider to be dual: the questioner must show that the answer has solved his problem (that he has marked the answer as the best answer) and that he Given a largely favorable mark (that it marked the answer as intrinsically good). The authors thus formalize the problem of prediction under the name of "problem of satisfaction of the questioner": to predict whether the questioner will be satisfied by the answers to his question.

The history of the CQA is rather short, it has already aroused a great deal of interest among researchers, ranging from information retrieval [25], resource comparison [26], Recommendation [27] To the user [28]. Current research on CQA services requires a study of the background, motivations and methods by which people seek and share their information. It may also involve the development of systems to support these activities.

Given the limited success of CQA's current automatic systems, another interesting way to solve a problem is to use crowd wisdom, also known as "collective intelligence". These social systems are called Community Question Answering (CQA).

CQA services generally consist of three elements: [29] first, a mechanism that allows users to submit their questions, secondly a complementary mechanism for users to provide answers to questions, and the third a flat -form web to facilitate user interactions. Online forums have acted as a CQA service function since the beginning of the Internet - so in this sense CQA is nothing new. CQA websites, however, have only appeared in recent years; The first CQA service, the Korean Naver Knowledge iN, was launched in 2002. The first CQA English website, Answer-bag, was launched in April 2003. CQA services have proliferated over the past eight years (If one considers the launch of Yahoo! Answers in 2005 as the milestone), as a rising market for the realization of various user intentions. It has been reported that the number of issues addressed in CQA services far exceeds the number of questions answered by the library reference services, [29] which was the main platform for answering these questions. In October 2009, Yahoo! Answers has more than 200 million users, of which more than 1.5 million users visit the site daily. In May 2010, it provided more than a billion questions, with on average an issue generated every 10 seconds; The number of questions submitted to China's CQA Baidu Knows service, to date, exceeded 155 million, with a daily volume of 10 million visitors.

Summary our proposition, by placing ourselves in the context of a CQA site, our goal is to provide a search engine with a relevant entry point for a query. Rather than displaying an entire page of answers of varying validity, it would be possible to discover the community's opinion and to find and display only the most relevant answer, possibly without even entering the CQA site. Of question answer, we thus consider the matching of the user's need for information to one or more relevant questions, is already achieved. The research problem arising from the objective previously formulated consists in estimating the best answer according to a question, all its answers and the metadata attached to it. This is reducible to a classification problem, with the "best" answers as a positive class, the rest as a negative class.

3. CLASSIFICATION APPROACH

3.1 Classification approach

Classification approach is a data analysis approach. It serves to facilitate the study and processing of data witlarge volumes [19]; this approach aims to consolidate data into groups to categorize data. It is a method among the methods used in data mining for the processing, analysis and exploitation of essential data. There are many classification approaches in data mining, namely neural networks, Bayesian networks and decision trees. The data is grouped into several classes such a way that the data of the same level are as similar as possible and the types are the most distinct possibility.

3.1 Decision trees or decision tree J48

History and operation; decision trees have their roots in Ross Quinlan's algorithm ID3 (for Iterative Dichotomiser 3) [31]. The goal of ID3 (and its improved versions: C4.5 and then J48) is the construction of a decision tree by maximizing Information Gain. The algorithm examines each attribute of the dataset, and determines each one has the enormous discriminant power (e.g., the most significant gain of information), iteratively, thus creating a tree structure.

Strengths and weaknesses; the power of decision trees lies in the intelligibility of the models that created. A model is directly understandable and interpretable by anyone, which is advantageous when one cannot blindly trust the model for ethical reasons (e.g., decision support in the medical field).

The decision tree is a data structure of statistical machine learning. Its operation based on heuristics that provide exceptional results in practice. Its appearance makes reading very clear and easy to exploit by humans [20].

The decision tree models a hierarchy of tests on the values of a set of variables called attributes. After these tests, the predictor produces a numerical value or selects an element in a discrete set of conclusions. We speak of regression in the first case and classification in the second [21].

There were many decision tree construction algorithms. Still, the most significant work was the CART algorithm (classification and regression trees) proposed by Breiman et al. In 1984 [22], the algorithm ID3 submitted by R. Quinlan in 1986 [23] and C4.5 algorithm which is an improvement of the algorithm ID3 also proposed by R. Quinlan in 1993 [24].

The Separator in Vast Marge

History and operation; The Vast Marge Separator (SVM, whose literal translation is Support Vector

Machine) was invented by Vladimir Vapnik and perfected by Vladimir Vapnik and Corinna Cortez in [13]. The SVM classifies the examples so as to maximize the distance between the hyper-plane separator and the examples closest to the separator hyper-plane (thus most likely to be erroneous), which gives it a robust character in the sense that The probability that the addition of new examples modifies the hyper-plane separator is minimized.

Strengths and weaknesses. The SVM is based on the theory of statistical learning and therefore benefits from assurances on the quality and robustness of its classification. It may, however, be slower than less theoretically based techniques such as decision trees or Bayesian decision.

4. CONTRIBUTION

4.1 Interest of our proposal

Our proposal is the realization of an automatic method to find the best entry point for an information retrieval system in a community question-answers-system (CQA). A vast amount of information, more than a billion answers for Yahoo! Answers only by [30], is available in CQA systems in a form that is only slightly exploited: the user seeking an answer to his question is obliged to go through all the answers to all Questions that are similar to his.

Finding the best entry point in a CQA system would allow a search engine to present answers to a user-generated question without even having to enter the system and without having to Complex reasoning on the part of the system.

This task requires predicting the answers that best satisfies the information requirement in the question.

4.2 Study of the Domain

4.2.1 Anatomy of an CQA System

CQA System is a set of questions, answers, agents and links between these three types of entities: an agent is the author of a question or answer, question. (Reputation, number of contributions,), questions (score, comments,) and answers (score, comments,). In the case of a simple navigation, the score of the questions is combined with a measure of freshness and diversity to present the questions to the user. In the case of a query, the score is combined with a measure of freshness and correspondence to the query to present to the user the best questions on the topic requested. Figure 1 illustrates the core of an CQA system.

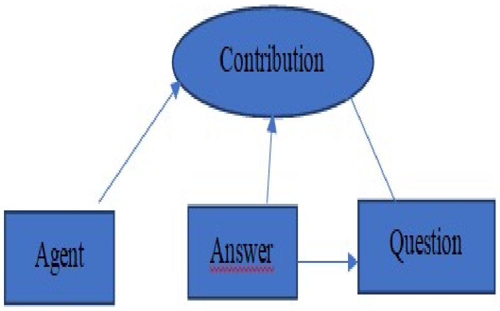


Figure1: illustrates CQA system

The organization of answers is based on their social score. The social score is defined as an aggregation measure of the community's opinion on the contribution. In his simplest incarnation, this consists in that:

- if an agent thinks the answer is false, he applies a negative vote and the social score of the answer is decreased by 1;

- if an agent thinks that the answer is correct, he / she applies a positive vote and the social score of the answer is increased by 1;

- in all other cases, the agent must restrict himself to vote.

This way of doing things both helps to raise the popularized answers by the community at the top of the page and to lower those that do not contribute to the resolution of the question. It can be complicated if necessary at the scale level (e.g., voting on a scale of ≠ 5 to +5, qualitative vote) or allocation of votes (e.g., egalitarian vote, random selection of voters) to correspond to Characteristics of the community. In addition to this, the author of the question has the ability to mark an answer as "valid", which means that the problem posed in the question has been solved using the content of the answer. The question is then considered resolved.

4.2.2 Formal Domain Model

Let Ω_Q be the universe of questions, Ω_R the universe of answers, Ω_A the universe of agents, and Ω_C the universe of contributions such that Ω_Q UΩ_R= Ω_C. A CQA system connects a set $R \subseteq \Omega_R$to each question, and a set of agents$A \subseteq \Omega_A$to their respective contributions to a time T. We also define the function f: $\Omega_R \times \Omega_R \times \Omega_A \rightarrow \mathbb{R}$ which associates a utility value with a Question Answer Agent triplet, which corresponds to the quality of the answers to the information requirement formulated by the question and to The epistemic state of the agent. This quality is between 0 and 1 inclusive. It should also be noted that for any set of answers there is an answers that includes this set. According to [14] and [15], a CQA system can satisfy three types of question:

- Factual issue,
- Request for opinion,
- Request for suggestion.

We will use an augmented typology based on this as well as that developed in [7].

- Factual information;
- General advice;
- Personal counseling;
- General opinion;
- Personal opinion;
- Other.

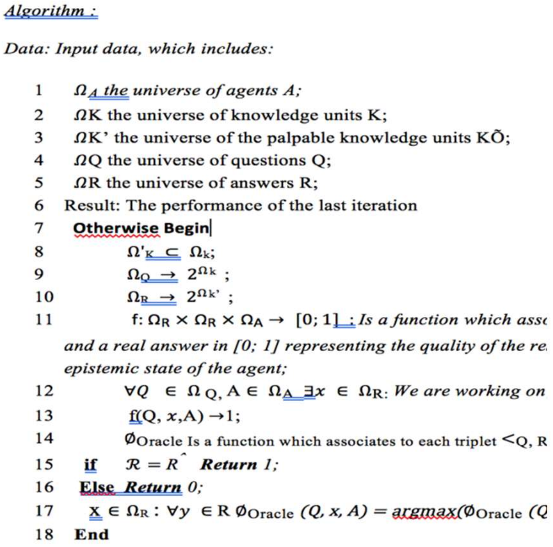
These four types of questions (other is not considered) can be differentiated by the need for information that they formulate.

A factual question corresponds to a need for fixed information, which can be one or more answers. This results in the existence of an answer R œ R associated with the question Q and the agent A such that f (R, Q, A) = 1. The satisfaction of the need for information depends mainly on the content of the Answer, and little of the agent. It should be noted that because for each set of answers there is a answers that includes each of its elements, the answers that fully meets the information need may not be in the proposed answers but be a composition of some parts of each reply. In this case, it will be assumed that the agent will select as the "best answer" the answer that is closest to the prototype answer.

A request for opinion (general or personal) corresponds to a need for variable information, which cannot be met by one or more answers. The questioner chooses arbitrarily when he has accumulated enough opinions or when an opinion suits him well, and the satisfaction of the need for information is not dependent on the content of the answers, but simply on their presence (Or absence). This results in the inequality$\forall \mathcal{R} \in \Omega_R$: f (R, Q, A) ≤1.

A request for advice (general or personal) corresponds to a need for varied information, it is possible but not guaranteed that there is a answers that satisfies it, and the satisfaction of the need for information depends mostly on the agent And little of the content of the answer. This results in the inequality $\forall \mathcal{R} \in \Omega_R$: f (R, Q, A) ≤ 1.

We can therefore formally define an AQ in the following way: Let:



```
Algorithm :

Data: Input data, which includes:

    $\Omega_A$ the universe of agents A;
    $\Omega$K the universe of knowledge units K;
    $\Omega$K' the universe of the palpable knowledge units KÕ;
    $\Omega$Q the universe of questions Q;
    $\Omega$R the universe of answers R;
  Result: The performance of the last iteration
    Otherwise Begin
            $\Omega'_K \subset \Omega_k$;
            $\Omega_Q \rightarrow 2^{\Omega k}$ ;
            $\Omega_R \rightarrow 2^{\Omega k'}$ ;
             f: $\Omega_R \times \Omega_R \times \Omega_A \rightarrow [0; 1]$ : Is a function which ass
  and a real answer in [0; 1] representing the quality of the re
  epistemic state of the agent;
            $\forall Q \in \Omega_Q, A \in \Omega_A \exists x \in \Omega_R$: We are working on
            f(Q, x,A) →1;
            ∅Oracle Is a function which associates to each triplet <Q, R
     if     $\mathcal{R} = \hat{R}$  Return 1;
    Else Return 0;
     $x \in \Omega_R : \forall y \in R$ ∅Oracle (Q, x, A) = argmax(∅Oracle (Q
   End
```

Figure2:The following algorithm is proposed to select and find the best answer in

the set of answers already available,

Our goal is to find the best answer in the set of answers already available, which means finding: $x \in \Omega_R$:$\forall y \in R$ ∅Oracle (Q, x, A) = argmax(∅Oracle (Q, y, A),

Ω_R is our decision space. It is different:

- $\hat{R}$ is the best solution in the subset of R that has been mapped;

- R* is the best solution in Ω_R.

It is possible that $\hat{R} \neq$ R*. Our goal is to find $\hat{R}^{12}$. Our formalization also makes it possible to express the "problem of satisfaction of the questioner" formulated by Liu et al. [5]: "According to the question submitted by a given agent in a CQA system, predict whether the agent will be satisfied with the community's answers." This is translatable as follows:

- Is CQA system;

- Predicting whether an agent is satisfied by the answers to his question is equivalent to doing the Non-exclusive disjunction of ∅Oracle (Q, x, A) → {0; 1} for each answer.

4.3 First approach: prediction of the best answer

In this section we detail our methodology for predicting the best answers in an CQA system.

4.3.1 Identification of Significant Attributes

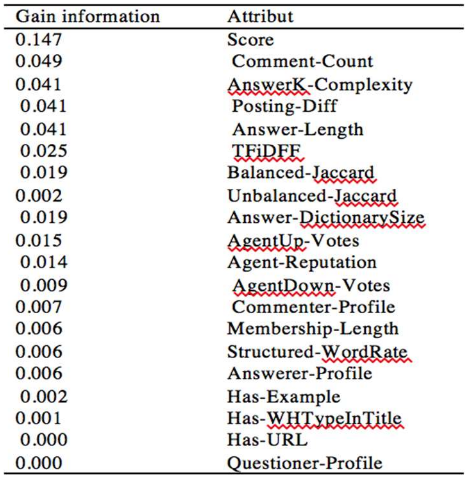
Description of Attributes

- Score: the social score of the answer;
- Has-URL: Does the answer contain a URL?
- Comment-Count: the number of comments of the answer;
- Answer K-Complexity: an approximation of the complexity of Kolmogorov of the text of the
- Answer, computed by the Lempel-Ziv algorithm LZ78 [32];
- Answer-Length: the length of the text of the answer;
- - TFiDFF: the similarity between the text of the question and that of the answer, calculated by theModel BM25 [4];
- Balanced-Jaccard: a similarity of Jaccard between the dictionary of the question and that of the answer, calculated by the formula J (A, B) = $\frac{A \cap B}{A \cup B}$;
- Has-Example: Does the answer have linguistic traces of an example?
- Unbalanced-Jaccard: a measure of how well the question is contained in the answer, calculated by the formula UJ (A, B) = $\frac{A \cap B}{B}$;
- Answer-Dictionary Size: dictionary size of the answer;
- Agent-Up Votes: number of positive votes received by the agent;
- Has-WH Type In Title: is the title of the question in WH- (e.g., "what", "why", etc.)?
- Agent-Reputation: agent's total reputation;
- Posting-Diff: the difference (in days) between the question and the answer;
- Agent Down-Votes: number of negative votes received by the agent;
- Comment-Profile: the proportion of former agent contributions that are comments;
- Answerer-Profile: the proportion of old agent contributions that are answers;
- Agent-Reputation: agent's total reputation;
- Membership-Length: the number of days the agent was a member of the site;
- Structured-Word Rate: ratio of structure words (e.g., "First", "because", etc.) in
- the answer ;
- Answerer-Profile: the proportion of old agent contributions that are answers;

- Answer-Dictionary Size: dictionary size of the answer;
- Agent Up-Votes: number of positive votes received by the agent;
- Questioner-Profile: the proportion of past agent contributions that are questions;
- Agent Down-Votes: number of negative votes received by the agent;

These attributes were chosen as easily recognizable manifestations of automated criteria that intuitively represent the quality of aanswer. An influence digraph can be found in the figure which illustrates the influence of each attribute on Each hidden variable.

Table1 . The 20 Attributes We Chose



Gain information	Attribut
0.147	Score
0.049	Comment-Count
0.041	AnswerK-Complexity
0.041	Posting-Diff
0.041	Answer-Length
0.025	TFiDFF
0.019	Balanced-Jaccard
0.002	Unbalanced-Jaccard
0.019	Answer-DictionarySize
0.015	AgentUp-Votes
0.014	Agent-Reputation
0.009	AgentDown-Votes
0.007	Commenter-Profile
0.006	Membership-Length
0.006	Structured-WordRate
0.006	Answerer-Profile
0.002	Has-Example
0.001	Has-WHTypeInTitle
0.000	Has-URL
0.000	Questioner-Profile

4.3.2 Experimental validation protocol

The evaluation of the classifiers is done by separating the data set into two parts: the training game, which the classifier uses to construct its model, and the test set, which the classifier uses to evaluate its model according to different measures.

We are looking for the best estimate of their true error, that is, the classification error rate they will have on all existing data. Since this error is not realistically calculable, we must approach it.

The closest approximation to the real error estimate is cross-validation in k-passes. Cross-validation in k passes consists first of all in splitting the data set into k partitions. A model is subsequently constructed at each of the k iterations on all partitions except one, which is retained as a test set. Performance is ultimately calculated by averaging performance on k models. 10-fold cross validation is the standard measure. It should be noted that at its extremes, cross-validation becomes a standard test set or becomes a validation called Leave-One-Out, or All-Except-one (with as many passes as There are examples).

4.3.3 Performance Measures

The performance measures used are related to the task we wish to accomplish. Classification performance depends on two contradictory objectives: limiting type 1 errors, or false positives, and limiting type 2 errors, or false negatives. Type 1 errors are manifested by noise: the user is presented with an answer which is considered good but which is not. Type 2 errors manifest themselves in silence: failure to present a good solution to the user. The aim is to minimize both noise and silence. The measures that characterize the competence of the system on these criteria are precision and recall.

- noise : false positives;

- silence : false negatives;

-Precision: $\frac{\text{true positives}}{\text{True positive + false positives}}$;

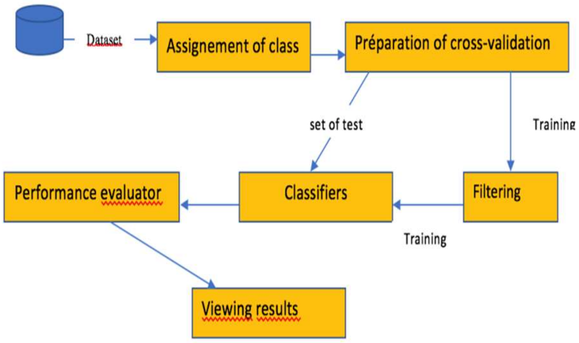


Figure 3 : The learning and evaluation process

-recall: $\frac{\text{true positives}}{\text{True positive + false positives}}$;

Considering the "best answers" as the positive class and the "no better answers" as the negative class, the precision corresponds to the proportion of positives returned by the classifier that belong to the positive class: it is therefore a measure of purity . The recall corresponds to the proportion of the positive class returned by the classifier: it is therefore a measure of completeness. Precision is opposed to noise and recall is opposed to silence.

Some measurements, such as the Fβ-measure, can be used to aggregate precision and recall. It is then possible to modify the parameter β to fix the influence of the precision or the recall on the score. The three most common versions are F_1-measure, where precision and recall have the same importance, F_2-measure, where recall is more important, and $F_{0.5}$-measure, where precision is more important.

Fβ-measure= $\frac{(1+\beta^2).(Accuracy + reminder)}{(\beta^2.\ Accuracy + reminder)}$

The goal is to help the user navigate by allowing the information retrieval systems to present a direct answerto the user's question, we propose to privilege the recall. The best-answer prediction is thus transformed, in the worst case, into information filtering.

4.4 Second approach: estimation of the weights of each parameter

4.4.1 Objectives and contextualization of symbolic regression to find a white box model

Symbolic regression allows us to estimate an equation using our data set to predict, with the minimum number of errors, the class of each answer from its attributes. To the "black box" model of automatic learning, symbolic regression substitutes a "white box" model.

Table 2: The 10 most discriminating attributes, calculated by the gain of information

Gain information	Attribut
0.147	Score
0.049	Comment-Count`
0.041	AnswerK-Complexity
0.041	Posting-Diff
0.041	Answer-Length
0.025	TFiDFF
0.019	Balanced-Jaccard
0.002	Unbalanced-Jaccard
0.019	Answer-DictionarySize
0.015	Agent-UpVote

4.4.2 Performance Measures

Symbolic regression uses genetic programming as a support. Genetic programming works by minimizing a fitness function that characterizes how well the current solution is adapted to its environment (the problem to be solved). This aptitude function can be calculated in different ways and used as a measure of performance, and is in most cases calculated as an inverse function of the error of the individual. The two most commonly used error functions are:

- the mean logarithmic error: $\frac{1}{N}\sum_{i=1}^{N} log(1+y-f(x))$is used when the dataset has noise with outliers;

Attribute

- The mean squared error: Has a normally distributed sound; $\frac{1}{N}\sum_{i=1}^{N}(y-f(x))^2$is used when the data set

Since we standardize and standardize our data, MSE is the most appropriate error function.

5. VALIDATION

5.1 Behavior and Habits of Stack Over flow Users

At this time, StackOverflow has 3.1 million questions, 6.2 million answers, 12 million comments. There are on average 2 answers per question, 4 comments per questions.

On a sample of 100,000 randomly extracted questions between 2008 and 2010:

- 31% do not have a "better answer";
- not interested.

5.2 Preparation of data

Data preparation consists of three phases: the elimination of external biases that could influence prediction, the elimination of the bias induced by the resampling of the data, and the processing of the attributes (standardization).

Elimination of temporal bias

When an agent selects a answers as the best answer, it does so with knowledge of other answers available at that time. The probability of choosing a reply is distributed among the answers of which he is aware (according to the axiom of the choice of Luce 17). If other answers arrive in the meantime, they are "disqualified" by default, by virtue of being late arrivals. For this reason, we eliminate the answers that were created after the best answer was chosen.

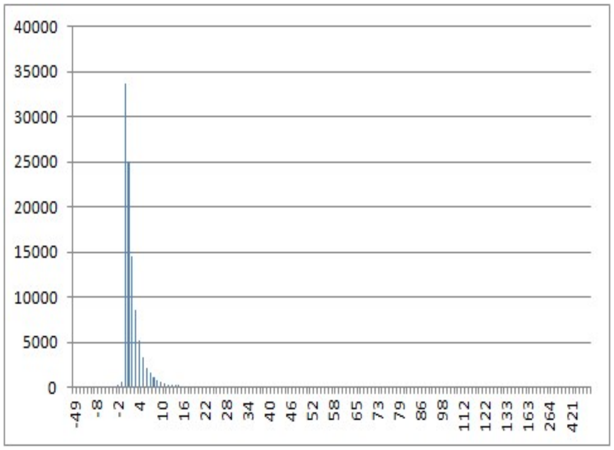


Figure 4:illustrating the distribution of scores questions answering

Preparing a benchmark

Our dataset, extracted from StackOverflow, is different from the usual data sets of literature. We consider that the prediction Yahoo! Answers use too many idiosyncratic features of these systems. Moreover, the generality of these sites gives a clear advantage to techniques based on the apparently lexical, typical of the search for information. StackOverflow is moderated with extreme vigilance to ensure that any discussion taking place on the site is entirely technical. The corpus of data we are working on is therefore neutral, technical and devoid of StackOverflow idiosyncrasies, which we believe are the criteria of a reasonable benchmark for entry point prediction in QRC systems. In order to have a comparative heel, is shown in bold on the scores details the best score obtained using the subset of extractable features on our dataset.

Table 3. Best Literatureattribute Score

Predicting Information Seeker Satisfaction [5]	Precision Recall	70.4% 65.0%
Evaluating and Predicting Answer Quality[12]	Precision Recall	71.0% 68.4%
Analyzing and Predicting Not-Answered [9]	Precision Recall	63.3% 65.0%
A Framework to Predict the Quality of Answers [3]	Precision Recall	69.0% 61.3%

Table 4.Results

Algorithm	Measure	Score
SVM	Precision Reminder F1-measure	57.1% 56.9% 55%
Decisiontree	Precision Reminder F1-measure	71.0% 71.0% 71.0%

5.3 Identification of the best answers

5.3.1 Definition of the best answers concept

We decide to adopt as a definition of best answer a common user feedback: the fact that the questioner has marked the question "best answer" or not. An answer is deemed "best answer" if the questioner has marked it as such. This allows us to retrieve a large amount of pre-labeled data.

5.3.2 Results of the experiment

We refer to

Confusion matrices

Interpretation of results

We see that the decision tree with Bagging has a higher precision and a stronger reminder. It is also possible to note the lack of e and Boosting. This is generally due to the fact that there is a lot of noise in the data, and suggests that in our future work we should find new attributes capable of counteracting the effect of noise.

Table 5.Results with Bagging

Algorithme	Measure	Score	
Decision tree	Precision Reminder F1 -measure	68.8% 68.8% 68.8%	Boosting
Decision tree	Precision Reminder F1 -measure	72.3% 72.2% 72.1%	Bagging

Table 6 .Confusion Matrix For The Decision Tree

Negative	Positive	→ Classified as
9816	4465	Negative
3640	**10641**	Positive

Table 7 .SVM confusion matrix

Negative	Positive	→Classified as
5935	8200	Negative
4140	**10185**	Positive

Table 8.Confusion Matrix For The Boosted Decision tree

Negative	Positive	→ Classified as
9607	4674	Negative
4189	**9993**	Positive

5.4 Symbolic regression

5.4.1 Preparation of data

The process of symbolic regression is an iterative process. It is necessary to explore the space of the states in a preliminary way to know how to adjust each of the parameters.

5.4.2 Definition of building blocks

The first phase consists of the definition of the building blocks. The building blocks will be the basic instructions that the genetic programming software will manipulate with the aim of approximating a theoretical function defined by the supplied data. Since we wish intelligible equations, it is necessary to take light instructions. Moreover, since we are looking for a predictive model, we need to use a form of cross validation on our equations, which the Eureqa Formulize software does independently

5.4.3 Results

The two best equations are described in the following enumeration.

- IsBest = step (sgn (ScoreAnswerKComplexity ≠ AnswerDictionarySizeand (Score, Score))) With a correlation coefficient of 0.79 and an MSE of 0.08

- IsBest = step (sgn (ScoreúAnswerKComplexity ≠ AnswerDictionarySizeúcos (Score))) With a correlation coefficient of 0.79 and an MSE of 0.08

Interpretation

If it is impossible to derive a multicriteria aggregation model from these equations, we can nevertheless see that the Score and the algorithmic complexity of the answersplay an important role, suggesting that a search path to a mathematical analysis of the complexity of a message could be viable.

Table 9.Confusion Matrix for Bagging Decision Tree

Negative	Positive	→ Classified as
9727	4464	Negative
3426	**10765**	Positive

5.4.4The results of the recommendation of a direct answer to the question formulated by the user

Table 10: Precision / Recall

Percentageof elimination	Measure	Precision	Recall
25%	Precision	0.9604	0.2501
	Reminder	0.7610	0.4812
	F1 -measure	0.5809	0.6324
50%	Precision	0.3416	0.1435
	Reminder	0.2700	0.1863
	F1 -measure	0.2031	0.2500
75%	Precision	0.1781	0.1166
	Reminder	0.1592	0.0265
	F1 -measure	0.1373	0.0085

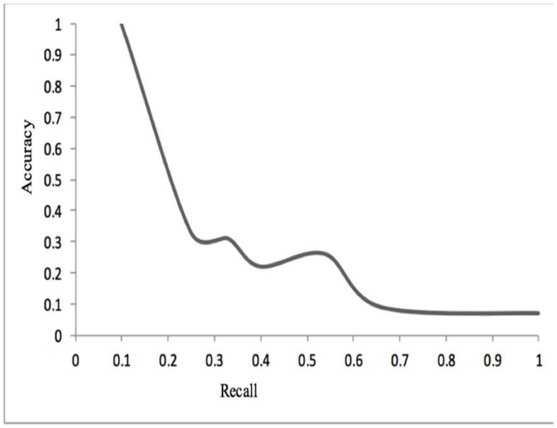


Figure 5: The precision / recall curve

We note that for low values of accuracy, recall becomes important and vice versa. We can select the optimum value for our system it depends on our need.

If the accuracy is low, the user will be dissatisfied, because he will have to waste time reading replies that are not of interest to him. If the reminder is weak, the user will not have access to a direct answer to the question formulated by the user he / she would like to have. A perfect information retrieval system must have a precision and a reminder close to the value 1, but these two requirements are often contradictory and a very high precision can be obtained only at the price of a weak recall and vice versa.

6. CONCLUSION

We have devoted ourselves to the study of CQA systems through their structure and the different meta-data they produce, with minimal use of the content, limited to the lexical ones. An interesting perspective would be to combine our approach with finer techniques of automatic language processing, in order to determine the quality of answers according to other criteria, such as intelligibility, writing style or the specificity of the terms used our prospects are focused on three areas of research:

1. continue the formalization in order to standardize it with the formalization of the information needs developed by Eduard Hoenkamp in [16]. The author formalizes the notion of need for information using the Galois lattice theory and uses it to make the link between the different forms of information retrieval;
2. to treat the notion of social signals (developed in this paper) as an aid to prediction and to combine it with a more in-depth analysis of answers, in the form of automatic language processing (for the content of the answer) and Analysis of social networks (for relations between agents);
3. explore the possibility of combining a top-down system that uses CQA systems as a knowledge base with a bottom-up system of knowledge inference.

REFERENCE:

[1] Jurczyk P. et Agichtein E. : Discovering authorities in question answer communities by using link analysis. in Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, CIKM ’07, p. 919–922, New York, NY, USA, 2007. ACM.

[2] Jurczyk P. et Agichtein E. : Hits on question answer portals: exploration of link analysis for author ranking. in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR ’07, p. 845–846, New York, NY, USA, 2007. ACM.

[3] Jeon J., Croft W. B., Lee J. H. et Park S. : A framework to predict the quality of answers with non-textual features. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in

information retrieval, SIGIR '06, p. 228–235, New York, NY, USA, 2006. ACM.

[4] Jones K. S., Walker S. et Robertson S. E. : A probabilistic model of information retrieval: development and comparative experiments. Inf. Process. Manage., 36(6): 779–808, nov. 2000.

[5] Liu Y., Bian J. et AgichteinE. : Predicting information seeker satisfaction in com- munity question answering. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '08, p. 483–490, New York, NY, USA, 2008. ACM.

[6] Liu C., Yeung C. H. et Zhang Z.-K. : Self-organization in social tagging systems. CoRR, abs/1102.3989, 2011.

[7] Mendes Rodrigues E. et Milic-FraylingN. : Socializing or knowledge sharing?: characterizing social intent in community question answering. In Proceedings of the 18th ACM conference on Information and knowledge management, CIKM '09, p. 1127–1136, New York, NY, USA, 2009. ACM.

[8] O'Reilly T. : What is web 2.0: Design patterns and business models for the next generation of software. MPRA Paper 4578, University Library of Munich, Germany, mars 2007.

[9] Yang L., Bao S., Lin Q., Wu X., Han D., Su Z. et Yu Y. : Analyzing and predicting not-answered questions in community-based question answering services.

[10] Golder S. A. et Huberman B. A. : The structure of collaborative tagging systems. CoRR, abs/cs/0508082, 2005.

[11] Kleinberg J. M. : Authoritative sources in a hyperlinked environment. J. ACM, 46(5):604–632, sep. 1999.

[12] Shah C. et PomerantzJ. : Evaluating and predicting answer quality in community CQA. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, SIGIR '10, p. 411–418, New York, NY, USA, 2010. ACM.

[13] Cortes C. et Vapnik V. : Support-vector networks. Machine Learning, 20:273–297, 1995. 10.1007/BF00994018.

[14] Kim S., Oh J. S. et Oh S. : Best-Answer Selection Criteria in a Social Q&A site from the User-Oriented Relevance Perspective. In ASIST, 2007.

[15] Harper F. M., Raban D., Rafaeli S. et Konstan J. A. : Predictors of answer quality in online q&a sites. In Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, CHI '08, p. 865–874, New York, NY, USA, 2008. ACM.

[16] HoenkampE. : On the notion of "an information need". In Proceedings of the 2nd International Conference on Theory of Information Retrieval: Advances in Information Retrieval Theory, ICTIR '09, p. 354–357, Berlin, Heidelberg, 2009. Springer- Verlag.

[17] Yandong Liu and Eugene Agichtein. On the evolution of the Yahoo! answers QA community. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pages 737–738, Singapore, 2008.

[18] Tamer Mohamed Elsayed. Identity resolution in email collections. PhD the- sis, University of Maryland at College Park, College Park, MD, USA, 2009. AAI3372840.

[19] Stéphane Tuffery. Data mining et statistique décisionnelle. Editions Technip, Août 2005.

[20] Laurent Hyafil and R. L. Rivest. Constructing Optimal Binary Decision Trees is NP-complete. Information Processing Letters, 5(1):15–17, 1976.

[21] Leo Breiman. Random forests. Machine Learning, 45(1):5–32, 2001.

[22] Leo Breiman, Jerome Friedman, Charles J. Stone, and R. A. Olshen. Classification and Regression Trees. Chapman and Hall/CRC, 1 edition, January 1984.

[23] J. Ross Quinlan. Induction of decision trees. Machine Learning, 1(1):81–106, 1986.

[24] J. Ross Quinlan. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[25] Yunjie Xu, Hee-Woong Kim, and AtreyiKankanhalli. Task and social information seeking: Whom do we prefer and whom do we approach? J. Manage. Inf. Syst., 27(3):211–240, January 2010.

[26] F. Maxwell Harper, Daphne Raban, SheizafRafaeli, and Joseph A. Konstan. Predictors of answer quality in online q&a sites. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '08, pages 865–874, New York, NY, USA, 2008. ACM.

[27] Maggy Anastasia Suryanto, Ee Peng Lim, Aixin Sun, and Roger H. L. Chiang. Quality-aware collaborative question answering: methods and evaluation. In Proceedings of the Second ACM International Conference on Web Search and Data Mining, WSDM '09, pages 142–151, New York, NY, USA, 2009. ACM

[28] F. Maxwell Harper, Daniel Moy, and Joseph A. Konstan. Facts or friends?: Distinguishing informational and conversational questions in social Q&A sites. In Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI), pages 759–768, Boston, MA, USA, 2009.

[29] Chirag Shah and JeffereyPomerantz. Evaluating and predicting answer quality in community CQA. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, SIGIR ’10, pages 411–418, New York, NY, USA, 2010. ACM.

[30] Team Y. A. : 1 billion answers served! http://yanswersblog.com/index.php/archives/2010/05/03/1-billion-answers-served/, 2010.

[31] Quinlan J. R. : Induction of decision trees. Machine Learning, 1(1):81–106, mars 1986.

[32] Ziv J. et Lempel A. : A universal algorithm for sequential data compression. IEEE TRANSACTIONS ON INFORMATION THEORY, 23(3):337–343, 1977.